

Automatic, optimized interface placement in forward flux sampling simulations

Kai Kratzer,¹ Axel Arnold,¹ and Rosalind J. Allen²

¹*ICP, Institute for Computational Physics, University of Stuttgart, Allmandring 3, 70569 Stuttgart, DE.*

²*SUPA, School of Physics, The University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK.*

(Dated: April 3, 2013)

Forward flux sampling (FFS) provides a convenient and efficient way to simulate rare events in equilibrium or non-equilibrium systems. FFS ratchets the system from an initial state to a final state via a series of interfaces in phase space. The efficiency of FFS depends sensitively on the positions of the interfaces. We present two alternative methods for placing interfaces automatically and adaptively in their optimal locations, on-the-fly as an FFS simulation progresses, without prior knowledge or user intervention. These methods allow the FFS simulation to advance efficiently through bottlenecks in phase space by placing more interfaces where the probability of advancement is lower. The methods are demonstrated both for a single-particle test problem and for the crystallization of Yukawa particles. By removing the need for manual interface placement, our methods both facilitate the setting up of FFS simulations and improve their performance, especially for rare events which involve complex trajectories through phase space, with many bottlenecks.

I. INTRODUCTION

Many important processes in nature can be described as rare events – i.e. events that happen rapidly but unpredictably, with long waiting times between occurrences. Examples of such processes range from large-scale problems such as electricity or computer network failures, to molecular level processes such as the formation of crystal nuclei or vapour bubbles in metastable liquids. Rare events are difficult to study, either in experiments or in computer simulations, because most of the observation time is spent waiting for the fluctuation-driven event to happen. In simulations, this problem can be overcome using rare event simulation techniques such as umbrella sampling [1, 2], Bennett-Chandler methods [2–4], transition path sampling [5–7], transition interface sampling [8–10], milestoning [11–13], nudged elastic band [14, 15], string methods [16, 17], weighted-ensemble methods [18], non-equilibrium umbrella sampling or forward flux sampling (or splitting)-type methods [19–28]. All of these methods aim to enhance the sampling in the region of phase space (or trajectory space) that corresponds to the rare event, while reducing the amount of time the simulation spends in the uninteresting phase space (or trajectory space) regions corresponding to the waiting times.

In this paper, we focus on the forward flux sampling (FFS) approach [22]. In FFS, one uses an order parameter to measure the progress of the system from the initial state towards the final state. The region of phase space between the initial and final states is partitioned by a series of interfaces defined by specific values of the order parameter. These interfaces are used to ratchet the system from the initial to the final state. Short trajectories are fired from the initial state; if these reach the first interface, they are used as starting points for further trajectories, which, if they reach the second interface, are used as starting points for further trajectories, etc. During this procedure, the fraction of trajectories which reach the next interface is monitored. The product of

these “success probabilities” over all interfaces, together with the flux of trajectories out of the initial state, gives the transition rate from initial to final state. Unbiased transition trajectories can be reconstructed from the collection of trajectories between interfaces. FFS provides a convenient way to simulate rare events in stochastic dynamical systems, because it is rather simple to implement and allows direct calculation of the transition rate. Importantly, FFS is suitable for both equilibrium and non-equilibrium systems (since it does not require *a priori* knowledge of the phase space density) [22]. Recent advances in FFS-type methods include the development of different algorithms for the trajectory-firing procedure [20], computation of phase-space densities as well as transition rates [29], analysis and optimization of the efficiency of the method [21, 22, 24, 25, 28], and the development of FFS-like methods for systems which are out of the stationary state [30, 31]. While FFS is of course not a panacea for all rare event problems [32, 33], it is widely and successfully used for a range of systems, some of which would be difficult or impossible to tackle with other methods. The validity of FFS has been extensively tested against brute force simulations and other rare event simulation methods for a range of problems [19, 22, 34–37].

In this paper, we focus on the placement of interfaces in FFS. The number of interfaces and their positions are important inputs in FFS, since poor interface placement can have strongly detrimental effects on the efficiency [21, 38]. If the interfaces are placed too far apart the probability of reaching the next interface will be very low, and much effort will be wasted firing trajectories that fail to progress. On the other hand, if the interfaces are too close together, trajectories will be highly correlated between successive interfaces so that little new information is gained at each interface.

Borrero and Escobedo [38, 39] have shown that, for a fixed number of interfaces, the efficiency of FFS simulations is optimized when the flux of trajectories between interfaces is equalized: i.e., for a fixed number of trial

trajectories per interface, the probability of reaching the next interface should be equal for all interfaces. This criterion allows optimal interface placement, if one has prior knowledge of how the success probabilities depend on the order parameter. Such knowledge is, however, not usually available. Borrero and Escobedo suggest beginning with non-optimized interfaces and using the constant flux criterion to iteratively improve the interface placement in successive FFS runs [38, 39]. While this is a good strategy for some problems, it is problematic for computationally expensive systems with high barriers. For these systems, FFS simulations with poorly chosen interface sets simply will not finish in a reasonable computational time. This forces the user to spend much effort on finding a reasonable initial interface set, by manual trial-and-error. Moreover, repeating the FFS simulations to obtain iteratively better interface sets is computationally expensive. To our knowledge, the only interface-based method that does not require *a priori* interface placement is adaptive multi-level splitting (AMS) [40]. In this method, successive interfaces are placed adaptively, based on the furthest point in order parameter space reached by previous trajectories. Practical implementation of AMS is, however, more complex than for standard FFS, because one needs to keep track of the histories of previous trajectories in order to determine the start points of new trajectories. This is likely to involve coding and storage overheads, particularly for large systems.

In this paper, we present two methods which allow optimal placement of interfaces in standard FFS simulations, on-the-fly, *without any prior knowledge*. We first use theoretical arguments to estimate the optimal range for the flux between interfaces, when the number of interfaces is not fixed (section II). In section III we present our algorithms and discuss the situations in which we expect each to be advantageous. We demonstrate the performance of both methods in section IV, first for a one-particle test problem and then for a computationally expensive rare event problem: crystallization in a system of Yukawa particles. Finally we present our conclusions in section V.

II. OPTIMIZATION PRINCIPLES FOR INTERFACE PLACEMENT IN FFS

In this section, we first briefly describe the “direct” FFS algorithm. We then review the work of Borrero and Escobedo which shows that, for a fixed number of interfaces, optimal efficiency requires equal fluxes between interfaces [38, 39]. Building on this work, we establish the optimal range for the transition probability between interfaces, in the case where the number of interfaces is not constrained. This optimal range will be used as input for the computational algorithms described in section III.

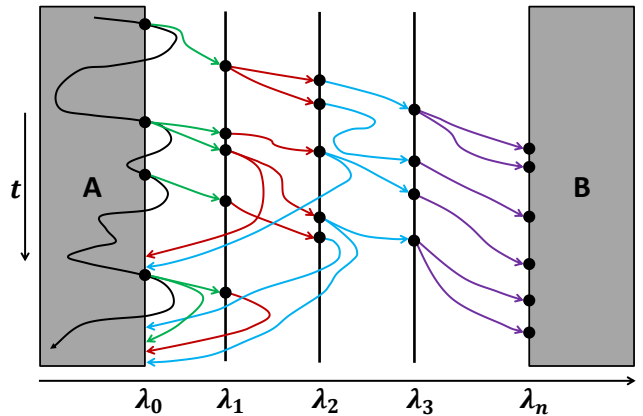


Figure 1: Schematic illustration of the DFFS algorithm. The barrier region between λ_A and λ_B is partitioned by a series of interfaces, defined as values of the order parameter λ (horizontal axis). The black dots denote stored configurations; and the coloured arrows (colour coded by interface) represent trajectories. The vertical axis denotes simulation time, increasing downwards.

A. The direct FFS algorithm (DFFS)

The aim of FFS is to compute the transition rate k_{AB} from an initial state A to a final state B , while at the same time sampling the associated transition trajectories. The transition rate k_{AB} is given by $k_{AB} = \Phi P_B$ [8], where Φ is the flux of trajectories leaving the initial state, and P_B is the probability that a trajectory that leaves the initial state will subsequently make it to the final state (rather than returning to the initial state). In FFS, the initial and final states are defined in terms of an order parameter λ , such that if $\lambda < \lambda_A$ the system is in the initial state and if $\lambda > \lambda_B$ it is in the final state. Intermediate values of λ ($\lambda_A < \lambda < \lambda_B$) correspond to the “barrier” region. This barrier region is partitioned by a series of n interfaces, defined by specific values of λ , such that $\lambda_i < \lambda_{i+1}$, $\lambda_0 \equiv \lambda_A$ and $\lambda_n \equiv \lambda_B$ (see Figure 1). The probability P_B can be written as [8]

$$P_B = \prod_{i=0}^{n-1} p_i \quad (1)$$

where p_i is the conditional probability that the system, having reached interface i , subsequently goes on to reach interface $i + 1$ before returning to the initial state. FFS provides a practical and efficient way to compute Φ , and p_i for each interface, thus allowing the computation of k_{AB} . The algorithm also generates transition trajectories. While several variants of FFS exist [20, 22, 28], we focus here on the direct FFS algorithm (DFFS) [19, 22, 28].

The DFFS algorithm has two stages. In the first stage, the flux Φ across the first interface λ_0 is computed by sim-

ulating a system in the initial state and monitoring the frequency with which the trajectory crosses λ_0 in the direction of increasing λ . When these crossings happen, the configuration of the system is stored; this simulation thus generates not only a measurement of Φ but also a collection of N_0 configurations corresponding to states of the system at the moments of crossing λ_0 [49]. In the second stage of the algorithm, the probabilities p_i are computed in a step-wise fashion (see Figure 1 for illustration). To compute p_1 , one chooses configurations at random from the collection stored at λ_0 and uses them to initiate new "trial" trajectories, which are continued until they either reach λ_1 ("success") or return to λ_0 ("failure"). For successful trajectories, the final configuration at λ_1 is stored in a new collection. After M_0 trial trajectories have been fired, p_1 is computed by dividing the number of successes by M_0 . One then repeats the same procedure, using the configurations at λ_1 as starting points for M_1 trajectories that are continued until they reach λ_2 or return to λ_0 , and so on, until the final interface is reached and one has a complete set of estimated probabilities p_i . Transition trajectories from the initial to the final state can then be reconstructed from the set of successful trajectories between interfaces [20, 22].

B. Equalization of fluxes between interfaces

Under the assumption that trajectories decorrelate between adjacent interfaces, analytic results for the computational efficiency of the DFFS method (and related methods) can be derived [21]. Even though these assumptions may not always be satisfied for many real problems, these analytical predictions still give a useful general guide to the performance of the method. In particular, by modelling the number of successful trajectories from interface i as a binomially distributed random variable with parameter p_i , one can obtain predictions for the computational cost, and the statistical error, associated with the computation of the rate constant k_{AB} for given choices of the number n of interfaces, the numbers M_i of trial trajectories, the number N_0 of configurations at λ_0 , and for given values of p_i [21]. For DFFS, the variance \mathcal{V} in the estimated rate constant is given approximately by [21, 22]

$$\mathcal{V} \approx \sum_{i=0}^{n-1} \frac{(1-p_i)}{M_i p_i}. \quad (2)$$

Borrero and Escobedo [28, 38] have shown that, for fixed n , $\{M_i\}$ and P_B , Eq. (2) can be minimized by placing the interfaces such that $M_i p_i$ is the same for all interfaces – i.e. the statistical error is smallest when the net flux of trajectories between successive interfaces is constant. Assuming, for simplicity, that one fires the same number of trajectories for each interface ($M_i = M = \text{const}$), one should place the interfaces such that p_i is the same for all interfaces. Thus, interfaces should be closer together

in "bottleneck" regions of the phase space (note that here Borrero and Escobedo assume that, since the number of interfaces is fixed, the computational cost does not depend strongly on the interface placement, and does not need to be considered in the optimization). An alternative formulation of the constant flux rule, put forward by Borrero and Escobedo, states that the quantity

$$f_i = \frac{\sum_{j=0}^{i-1} \log p_j}{\sum_{j=0}^{n-1} \log p_j} \quad (3)$$

should be linear when plotted against the interface index i . To see this we note that if all the transition probabilities are equal, $p_j = p = \text{const}$, then

$$f_i = \frac{\sum_{j=0}^{i-1} \log p}{\sum_{j=0}^{n-1} \log p} = \frac{i}{n}. \quad (4)$$

One can therefore measure the "quality" of a particular set of interfaces, either by directly asking whether the success probability p_i is the same across different interfaces, or by testing whether f_i is linear when plotted against the interface number i .

C. Optimal transition probability

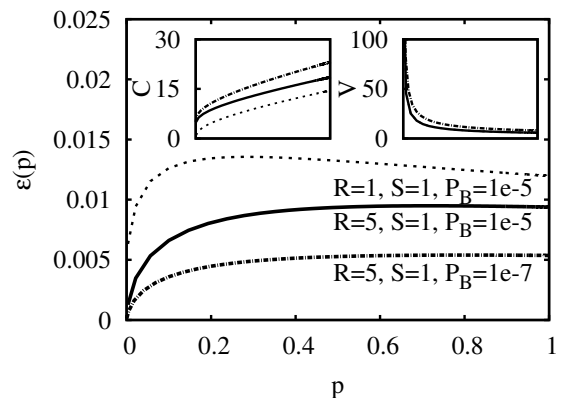


Figure 2: Theoretical prediction (Eq.(A9)) for the efficiency \mathcal{E} of a hypothetical rare event problem with $M = 200$ and $N_0 = 100$, plotted as a function of p for several values of P_B and of the cost parameters R and S (see text and Appendix A). The insets show the predicted computational cost \mathcal{C} (Eq.(A6)) and variance in the rate constant \mathcal{V} (Eq.(A8), which follows from Eq.(2)).

Borrero and Escobedo's work shows that for n interfaces, the optimal positioning is such that $p_i = p = P_B^{1/n}$ (this follows from Eq. (1)). However, if we are to place interfaces optimally, on-the-fly, we also wish to optimize the number n of interfaces. This is equivalent to optimizing the crossing probability p , under the constraint that $n = \log P_B / \log p$. Here we compute the optimal value of

1. Choose a trial position λ_{trial} for the next interface λ_{i+1} , in the range $\lambda_i < \lambda_{\text{trial}} < \lambda_B$. This should be done in a way appropriate to the problem being studied; we typically set $\lambda_{\text{trial}} = \lambda_i + b \times (\lambda_B - \lambda_A)$, where $0.01 < b < 0.1$, but one could also use for example $\lambda_{\text{trial}} = \lambda_i + (\lambda_i - \lambda_{i-1})$.
2. Using as starting points the configurations stored at λ_i , fire M_{trial} trajectories, which are continued until they reach λ_{trial} or λ_A . M_{trial} should be significantly smaller than the typical number of M trajectories per interface in the complete FFS simulation.
3. Compute $p_{\text{est}} = N_S / M_{\text{trial}}$ where N_S is the number of trial trajectories that reached λ_{trial} .
4. If $p_{\text{min}} < p_{\text{est}} < p_{\text{max}}$, accept the trial interface. Otherwise, choose a new trial interface position according to

$$\lambda_{\text{trial, new}} = \lambda_{\text{trial, old}} + \lambda_{\text{step}} \Delta p \quad (6)$$

where

$$\Delta p = \begin{cases} (p_{\text{est}} - p_{\text{max}}) & \text{if } p_{\text{est}} > p_{\text{max}} \\ (p_{\text{est}} - p_{\text{min}}) & \text{if } p_{\text{est}} < p_{\text{min}} \end{cases} \quad (7)$$

and fire trial trajectories to obtain a new p_{est} for this trial interface. Repeat this procedure until p_{est} lies within the desired range. If the resulting value $\lambda_{\text{trial, new}} < \lambda_i + d_{\text{min}}$, set $\lambda_{\text{trial}} = \lambda_i + d_{\text{min}}$, where d_{min} is the user-defined minimal acceptable distance between interfaces. If $\lambda_{\text{trial, new}} > \lambda_B$ set $\lambda_{\text{trial}} = \lambda_B$.

5. Set $\lambda_{i+1} = \lambda_{\text{trial}}$.
6. Continue with the DFFS simulation – i.e. fire M trajectories to λ_{i+1} to compute p_i and obtain a new collection of configurations at λ_{i+1} , as in the standard DFFS procedure. Any trajectories previously fired to this interface during step 5 can be included in the estimate of p_i .

In this algorithm, in addition to p_{min} and p_{max} , the user-defined parameters are the stepwidth λ_{step} (which determines how far the interface is shifted in each adjustment step), M_{trial} , the number of trial trajectories used to obtain p_{est} (typically $M_{\text{trial}} \approx 15$), and d_{min} , the minimal acceptable spacing between interfaces. This latter parameter is introduced to avoid excessive correlation between the sampling at successive interfaces, even if p_{min} is chosen to be small. The choice of d_{min} depends on the choice of order parameter and the dynamics of the system being studied. For example, if the order parameter is discrete, d_{min} should be at least one. In continuous systems, it should prevent the system from being able to cross several interfaces in a single timestep, and should be larger for systems whose dynamics is slow to decorrelate.

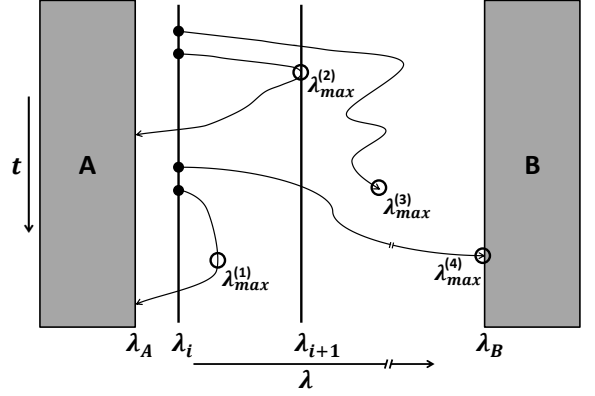


Figure 4: Schematic illustration of the exploring scouts method. A pre-defined number of trial trajectories are launched from the current interface λ_i . These trajectories continue until they reach the initial state A or the final state B, or until the maximum number of steps is reached. The maximum values of λ reached by the trial trajectories are then used to determine the position λ_{i+1} of the next interface.

The choice of interface shifting rule (point 4 in the algorithm described above) is not unique. We expect this rule to work well for systems with steep energy barriers, where one needs the initial interfaces to be closely spaced. However, for systems with flatter barriers, one might prefer to use a bisectional scheme, in which the trial interface is initially placed midway between λ_i and λ_B , and is then shifted forwards or backwards by bisecting the space between itself and either λ_i or λ_B .

The trial interface method is conceptually simple and can be implemented with only very minor modifications to an existing DFFS simulation code. The method also has the advantages that estimated transition probabilities for several possible trial interface positions can be computed in parallel on separate processors, and that any trial trajectories fired to interfaces that are eventually accepted can be reused in the final calculation of p_i . The method does, however, have the potential drawback that it relies on a reasonably good first estimate of λ_{trial} : if this first estimate is very poor, the algorithm may take many iterations to find an acceptable interface position. This problem is avoided in our second approach, the “exploring scouts” method.

B. Exploring scouts method

In the “exploring scouts” method, we again fire M_{trial} trial trajectories from interface λ_i , but this time without defining a trial interface position. In this method, illustrated in Figure 4, the trial trajectories are continued until they reach either λ_B or λ_A , or until a user-defined maximum number of steps is exceeded. The maximum value of λ achieved by each trial trajectory is monitored,

and the distribution of these values is used to position the next interface such that the success probability is close to a user-defined desired value p_{des} . The exploring scouts algorithm proceeds as follows:

1. Fire M_{trial} trial trajectories from interface λ_i , starting from the configurations generated by the DFFS algorithm. Continue each trajectory until it either reaches λ_B or λ_A , or exceeds m_{max} steps. Record the maximum value of λ achieved in each trial trajectory.
2. Generate a ranked list of maximum λ values for all trial trajectories – i.e. assign each trajectory an index k in the range $0 < k < M_{\text{trial}}$, such that $\lambda_{\text{max}}^{(k)} < \lambda_{\text{max}}^{(k+1)}$.
3. Compute $k_{\text{des}} = \lfloor M_{\text{trial}}(1 - p_{\text{des}}) \rfloor$ and set the position of the next interface $\lambda_{i+1} = \lambda_{\text{max}}^{(k_{\text{des}})}$. If the resulting value $\lambda_{i+1} < \lambda_i + d_{\text{min}}$, set $\lambda_{i+1} = \lambda_i + d_{\text{min}}$ (where d_{min} is the minimal acceptable spacing between interfaces as in the trial interface method).
4. Continue with the DFFS simulation – i.e. fire M trajectories to λ_{i+1} to compute p_i and obtain a new collection of configurations at λ_{i+1} , as in the standard DFFS procedure.

This algorithm works because the trial trajectories, or “exploring scouts”, supply information on the probability of reaching a particular value of λ , for all λ in the range $\lambda_i \rightarrow \lambda_B$. For entry k in our ranked list, k exploring scouts failed to reach $\lambda_{\text{max}}^{(k)}$ and $M_{\text{trial}} - k$ scouts reached $\lambda_{\text{max}}^{(k)}$ or beyond (note k runs from zero to $M_{\text{trial}} - 1$). The transition probability for an interface placed at $\lambda_{\text{max}}^{(k)}$ would therefore be approximately $(M_{\text{trial}} - k)/M_{\text{trial}}$. We can obtain a next interface position λ_{i+1} corresponding approximately to our desired transition probability p_{des} simply by picking the $\lfloor M_{\text{trial}}(1 - p_{\text{des}}) \rfloor$ -th entry in our list of maximal λ values. More precise versions of this algorithm are of course possible (e.g. interpolating between $\lambda_{\text{max}}^{(k)}$ values in our list). However, because the efficiency is in general not very sensitive to the precise value of p , we do not find these to be necessary.

The user-defined parameters for this method are the target probability p_{des} , the number M_{trial} of exploring scouts, the minimal interface spacing d_{min} and the limit m_{max} on the number of simulation steps per trial trajectory. If m_{max} is set too low, the algorithm will fail to explore regions of larger λ , and may tend to place the interfaces too close together (i.e. the true p will be smaller than p_{des}). Choosing a large value of m_{max} will, however make the algorithm more computationally expensive.

The exploring scouts method has the advantage that one knows *a priori* how many trial trajectories will be required to set the next interface position – this may be important in parallelized FFS applications. Furthermore, the number of user-defined parameters is fewer than in the trial interface method. The exploring scouts method

requires slightly more modifications to an existing standard DFFS code than the trial interface method, since one needs to track the maximal values of λ for the trial trajectories, but it is nevertheless rather simple to implement.

IV. EXAMPLES

We now demonstrate our interface placement methods for two test problems. First, we study the toy problem of a single particle undergoing Langevin dynamics in a one-dimensional potential; this also provides an opportunity to test the predictions for the computational efficiency made in section II C. Next, we demonstrate the utility of the methods for the much more computationally demanding example of crystal nucleation in a system of particles interacting *via* a Yukawa potential.

A. A single particle in a one-dimensional potential

We first consider a single particle moving in one dimension, in a potential with two minima, defined by $V(x) = (h/2)[1 - \cos(\pi x)]$ for x in the range $[-1, 3]$. The height of the potential barrier, at $x = 1$, is $h = 12k_B T$. The particle, which is initially placed in the region $-0.2 < x < 0.2$, undergoes underdamped Langevin dynamics. We set $k_B T = 1$, $m = 1$, $dt = 0.001$ and the friction coefficient $\gamma = 1$; with these parameters the crossover between ballistic and diffusive motion occurs on a timescale of about 1000 time steps or a dimensionless distance of 1. Our reaction coordinate λ is taken to be the position x of the particle and the borders of the initial and final states are defined by $\lambda_A = 0.2$ and $\lambda_B = 2$.

We carry out DFFS simulations for this problem, using both the trial interface method and the exploring scouts method. For both methods, we set $M_{\text{trial}} = 100$, $M = 1000$, $N_0 = 3000$ and $d_{\text{min}} = 0.01$. In the trial interface method, we set $p_{\text{min}} = 0.4$, $p_{\text{max}} = 0.6$ and the initial trial position for interface λ_{i+1} is chosen to be $\lambda_i + 0.1(\lambda_B - \lambda_i)$. In the exploring scouts method, we set $p_{\text{des}} = 0.5$ and $m_{\text{max}} = 10^5$.

Figure 5 shows the positions of the interfaces (main plot), and the resulting success probabilities p_i (inset), for the trial interface and exploring scouts methods. Both methods produce probabilities p_i that are approximately uniform across the interfaces, as desired. This corresponds to a highly non-uniform interface spacing: in fact all the interfaces are located prior to the maximum of the potential barrier, with the final interface lying close to the maximum. Figure 6 shows the quantity f_i , defined in Eq.(3), for both methods. This is indeed close to linear, confirming that the interface placement is close to optimal.

These methods allow us to choose the success probability p and place interfaces accordingly. We can there-

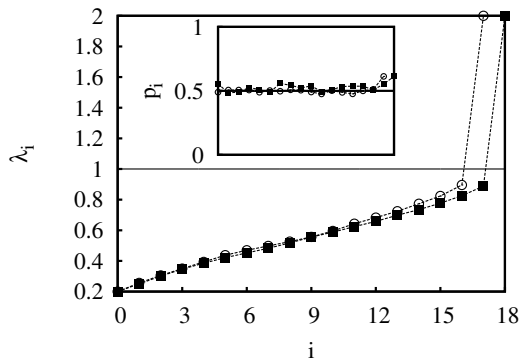


Figure 5: Single particle test case: positions λ_i of the interfaces as a function of interface index i , for the trial interface method (squares) and the exploring scouts method (circles). The maximum of the potential barrier is at $\lambda = 1.0$, shown by the solid horizontal line. The inset shows the success probabilities p_i , plotted as a function of interface index i .

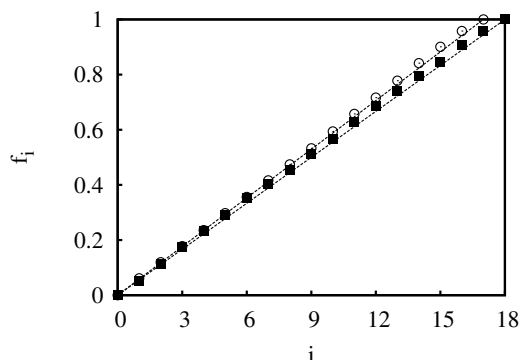


Figure 6: Single particle test case: f_i , as defined in Eq. (3), as a function of interface index i , for the trial interface method (squares) and the exploring scouts method (circles). The dashed lines show the optimal case where $f_i = i/n$ (see Eq. (4)). Note, that the two methods give slightly different numbers of interfaces.

fore use them to test the theoretical predictions made in section II C for the dependence of the computational efficiency on p . To this end, we have used the exploring scouts method to carry out a series of DFFS simulations for the single particle test problem, with the transition probability p varying between 0.05 and 0.95. The parameters of the method were as above, but with $M = 3000$. In these simulations, we measured the computational cost (in simulation steps) and the statistical error in the computed rate constant, allowing us to compute the computational efficiency \mathcal{E} as defined by Eq.(5). Figure 7 shows the measured computational efficiency, as a function of the transition probability p , compared to the theoretical prediction. The latter was computed using Eq.(A9), with $P_B = 1.36 \times 10^{-5}$ (the result obtained from our simulations), $R = 1.60 \times 10^7$ and $S = 1.39 \times 10^7$ (both in simulation steps, and obtained by fitting the cost function (A6) to our simulation data). The simulations are

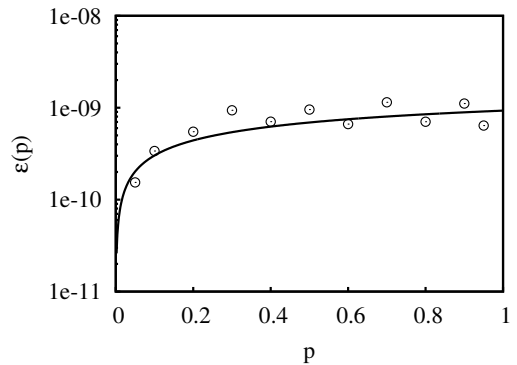


Figure 7: Single particle test case: computational efficiency \mathcal{E} as defined by Eq.(5), as a function of the success probability p . The data points show results of simulations using the trial interface method (for parameters see main text). The solid line shows the theoretical prediction of Eq. (A9) with $P_B = 1.36 \times 10^{-5}$, $R = 1.60 \times 10^7$ and $S = 1.39 \times 10^7$ (see main text). The number of interfaces placed by the algorithm varied between 3 (for $p = 0.05$) and 671 (for $p = 0.95$).

in remarkably good agreement with our theoretical predictions, showing that the estimated optimal values of p obtained from the theory are indeed valid, at least for this problem. Taking the error bars into account, the value of P_B obtained in the our simulations is independent of p , justifying the use of p as a performance tuning parameter and showing that the FFS method remains valid regardless of the number of interfaces (which varies between 3 and 671 as p is varied between 0.05 and 0.95).

B. Crystallization of Yukawa particles

We now move on to a much more challenging test problem: crystal nucleation in a system of particles interacting *via* a combined Yukawa and Weeks-Chandler-Andersen (WCA) potential [41], $U(r) = U_{\text{Yukawa}}(r) + U_{\text{WCA}}(r)$ with

$$U_{\text{Yukawa}}(r) = \epsilon \frac{\exp(-\kappa(r/\sigma - 1))}{r/\sigma} \quad (8)$$

and

$$U_{\text{WCA}}(r) = \begin{cases} 4 \left(\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 + \frac{1}{4} \right) & r < \sigma^{\frac{1}{6}} \\ 0 & \text{else.} \end{cases} \quad (9)$$

The repulsive WCA potential is used to model the excluded volume of the particles (note that the energy scale for the WCA potential is set to $k_B T = 1$ in our simulation units). The Yukawa potential is a screened Coulomb potential, suitable for modeling charged particles whose electrostatic interactions are screened by surrounding ionic atmospheres. In this work, the parameters of the Yukawa potential are the value of the repulsive potential at contact $\epsilon = 8$ (in units of $k_B T$) and the inverse

screening length $\kappa = 5$ (in terms of the hard-sphere diameter σ). Despite important previous advances [42, 43], the mechanism by which crystal nucleation happens in screened Coulomb systems remains an open question, to which FFS simulations can contribute by providing both nucleation rates and transition paths [42]. However, because the Yukawa interaction requires a larger cutoff radius than the more widely studied Lennard-Jones interaction, simulations of Yukawa particles are computationally expensive (especially for low salt conditions), which means that the number of trial trajectories which can be performed in an FFS simulation is limited. This makes setting up standard FFS simulations difficult, particularly under interesting conditions, e.g. close to coexistence where the transition rate is expected to be low [44]. Under these conditions, manual placing of the interfaces can easily lead to conditions where no FFS trial trajectories succeed in reaching the next interface. For such systems, automatic, optimal interface placement has the potential greatly to improve the feasibility and computational efficiency of FFS simulations.

We performed molecular dynamics (MD) simulations of 4096 WCA-Yukawa particles in a cubic box with periodic boundary conditions in the NPT ensemble at constant pressure $P = 38$ (LJ units) with a Langevin thermostat using the software package ESPResSo [45] in combination with DFFS, implemented in our rare event sampling framework FRESHS [46]. Note that FFS requires stochastic dynamics: here this is provided by the Langevin thermostat. The system is initially prepared in the liquid phase, which is undercooled (and therefore metastable). We are interested in the transition to the stable FCC crystal phase. Our order parameter λ is the size of the largest cluster of solid particles, where particles are classified as solid or liquid based on the local q_6 order parameter, as used in previous work [47]. The boundaries of the initial and final states were fixed such that the system is in the initial state if less than 0.5% of the particles are in the largest solid cluster and in the final state if more than 90% of the system's particles are in the largest solid cluster. This corresponds to $\lambda_A = 15$ and $\lambda_B = 3700$.

In our DFFS simulations, we compared three methods for interface placement: (i) placing the interfaces manually *via* a logarithmic scheme, (ii) the trial interface method and (iii) the exploring scouts method. All our DFFS simulations used $N_0 = 80$ configurations at the first interface and $M = 50$ trial runs per interface. Here, we discuss only the performance of the interface placement methods; the nucleation rates and pathways generated in the simulations will be presented elsewhere [48].

We first discuss the manual interface placement. For nucleation problems, where simulations are computationally very expensive, manual interface placement in FFS is very challenging. Our problem has a steep free energy barrier and so placing interfaces evenly between λ_A and λ_B results in very poor success rates for early interfaces. In fact, for our problem, we did not obtain any

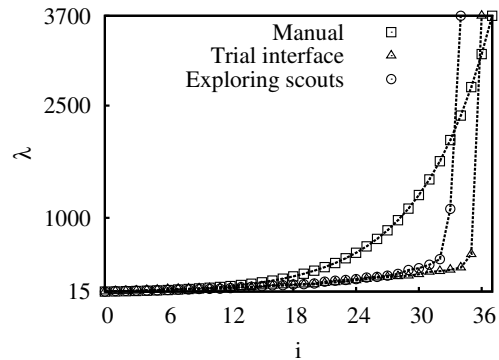


Figure 8: Yukawa test case: interface positions λ_i (plotted as a function of interface index i) generated by the manual interface placement (open squares), the trial interface method (open triangles) and the exploring scouts method (circles).

successes for early interfaces even with 100 evenly spaced interfaces. Therefore, as a “best possible” manual choice, based on this prior knowledge, we placed 36 interfaces logarithmically between λ_A and λ_B , with closer spacing between the early interfaces. Even with this rather well-informed choice of interfaces, Figure 9(a) shows that we obtain success probabilities that are far from equal across interfaces (inset). Indeed, many of the p_i values are very low: this is because the free energy landscape contains unforeseen bottleneck regions, in which too few interfaces were placed. Because the success probabilities are low in these bottleneck regions, much computational effort will be wasted on failed trajectories. Another problem is also apparent: for later interfaces, the transition probabilities are extremely high (close to 1). In this region of the free energy landscape, the crystal grows spontaneously: the placement of unnecessary interfaces implies extra computational overhead in storing configurations, etc. The fact that the manual interface placement is far from optimal is also apparent in the highly non-linear form of the function f_i when plotted against the interface index i (main plot in Fig. 9(a)).

We note that a commonly used approach to manual interface placement in FFS is to start with some initial guess, then if one obtains no successes for a given interface, shift it to a lower λ value and continue the FFS simulation. If not done carefully, this can actually bias the resulting computation of the rate constant k_{AB} towards higher values, since for interfaces at which by chance one obtains a large number of successes, one makes no change, but for interfaces where by chance one obtains few successes one shifts the interface. If such a shifting approach is used, bias can only be avoided by repeating the entire FFS simulation *a posteriori* - i.e. after the interface positions have been fixed.

Figures 8 and 9 also show the results of the automatic interface placement methods. For both methods, we set $d_{\min} = 3$ and $M_{\text{trial}} = 8$. The value of d_{\min} was chosen to prevent the system from crossing several interfaces in one

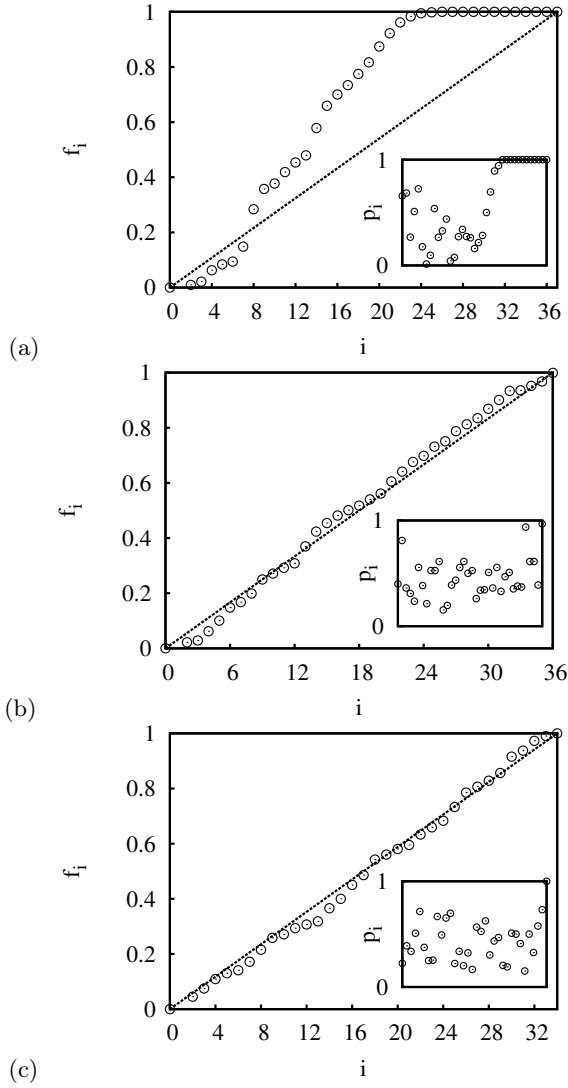


Figure 9: Yukawa test case: optimization criteria for the interface sets generated manually (a), with the trial interface method (b) and with the exploring scouts method (c) (for parameter values see main text). The main plots show the function f_i of Eq. (3), plotted against the interface index i , from our simulations (symbols) – the dashed lines show the optimal case, Eq. (4). The insets show the success probabilities p_i plotted against the interface index i .

MD timestep (simultaneous attachment of 3 particles in one step is unlikely) and to avoid correlation between trajectories at successive interfaces. For the trial interface method, we used $p_{\min} = 0.3$ and $p_{\max} = 0.6$ and the initial trial position for λ_{i+1} was set at $\lambda_i + 0.1(\lambda_B - \lambda_A)$. For the exploring scouts method, we used $p_{\text{des}} = 0.45$ and $m_{\max} = 10000$ timesteps. Figure 8 shows that both these methods produce similar interface numbers and positions, which are very different from those of our manual interface placement. The automatic methods position the interfaces much closer to the A state: in fact there are no interfaces at all for λ values greater than 1120.

Method	k_{AB}
Manual placing	$1 \times 10^{-14 \pm 2}$
Trial interface	$6 \times 10^{-14 \pm 1}$
Exploring scouts	$2 \times 10^{-14 \pm 1}$

Table I: Yukawa test case: rate constant k_{AB} (in $\sigma^{-3}\tau^{-1}$ with the simulation time unit τ) for DFFS simulations using the manual interface placement, trial interface method and exploring scouts method. For parameter values, see main text. The error bars in k_{AB} were determined by repeated independent simulations.

Method	Cost \mathcal{C}	Variance \mathcal{V}	Efficiency \mathcal{E}
Manual placing	7×10^6	6648	10^{-11}
Trial interface	4×10^6	188	10^{-9}
Exploring scouts	3×10^6	251	10^{-9}

Table II: Yukawa test case: computational cost, variance in the rate constant and resulting computational efficiency, for DFFS simulations using the manual interface placement, trial interface method and exploring scouts method. For parameter values, see main text. The cost was measured in simulation timesteps including the cost of exploratory trial runs for the automatic interface placement methods. The variance in the rate constant was estimated using Eq. (2), using simulation data for the p_i values.

This suggests that the free energy barrier to nucleation is located closer to λ_A than to λ_B – once the system has passed the barrier, the transition probability is always greater than the target value and thus no further interfaces are necessary. However, without *a priori* knowledge, there would be no way to guess this when placing the interfaces manually. Figure 9 (b) and (c) show that indeed both automatic interface placement methods perform well according to our optimization criteria: the success probabilities p_i are much more uniform, with no very low p_i values (insets). The functions f_i are also much more linear for the automatic interface placement methods than for the manual interface placement (main plots).

An obvious advantage to using the automatic interface placement methods is that setting up a DFFS simulation becomes very much easier and less time-consuming than using manual interface placement. In addition, the resulting DFFS calculations are more efficient with the optimized interface sets. Table I shows that the rate constants computed using the three interface placement methods are equivalent, but the error bars (computed by repeated FFS calculations) are larger for the manual interface placement. Moreover, as shown in Table II, the computational cost of the FFS calculation, measured in simulation steps, was about a factor of 2 lower for the automatically placed interfaces than for those that were placed manually. Had we not used our prior knowledge

to place the manual interface set logarithmically, this factor would have been even greater. For this problem, the exploring scouts method required about 25% fewer simulation steps than the trial interface method. Table II also shows estimates for the statistical error in the rate constant (computed using Eq. (2)), and the resulting computational efficiency. The estimated computational efficiency is two orders of magnitude higher using the automatic interface placement methods, compared to the manual interface set.

V. DISCUSSION

The efficiency of interface-based rare event simulation methods such as FFS is strongly dependent on the locations of the interfaces. Without *a priori* information, manual interface placement is a “hit and miss” task, that, for computationally intensive systems, often involves a large amount of user effort and results in non-optimal interface sets, for which the FFS calculations may be inefficient. In this paper, we have presented two methods for automatically placing interfaces on-the-fly in DFFS simulations, so that the user need only choose the order parameter, the definitions of the initial and final states, and the target transition probability (or its range). Building on previous work by Borrero and Escobedo, we have analysed theoretically how the computational efficiency depends on the interface transition probability p , providing an analytical expression for the optimal value of p for a given total transition probability P_B . We have further shown that in fact this optimum is broad and not very sensitive to P_B , so that for most problems target success probabilities in the range 0.3 – 0.7 are likely to produce satisfactory results. The lower bound of this range is set by the fact that efficiency decreases strongly when the success probability becomes too low. The upper bound is determined by the fact that trajectories will be highly correlated at successive interfaces if they are too close, meaning that little extra information is gained.

Our two methods for automatic interface placement both work by firing a small number of trial trajectories from an existing interface, to determine the position of the next interface. The methods differ in the way in which the information from these trial trajectories is used. In the trial interface method, a “trial” interface is placed, the probability of reaching this interface is estimated, and the trial interface is shifted until the estimated probability lies within an acceptable range. This method is very simple to implement in an existing DFFS code, because the information needed from the trial runs (simply whether they succeeded or failed) is the same as in a conventional FFS simulation. The interface shifting step can easily be parallelized and information from some of the trial runs can be re-used in the actual FFS step once the interface has been fixed. The exploring scouts method is in some ways more sophisticated: here, trial runs are fired from the existing interface and the dis-

tribution of the maximum λ values which they reach is used to determine a position for the next interface which corresponds to the target probability. This method has the advantage that one knows *a priori* how many trial runs will be needed to fix the interface position (important in some parallel implementations of FFS) and that the maximum length of these trial runs is fixed (albeit with some loss of accuracy in the interface position if the runs are too short). This may be important for problems where trial runs require many computational steps (e.g. free energy barriers that are not sharply peaked, or where returning to the initial state happens slowly). However, implementation of the exploring scouts method requires slightly more modifications to existing DFFS codes, since one needs to know the maximal value of λ reached by the trial runs, rather than their success or failure, as in standard DFFS. While we did not test it here, one could of course combine the trial interface and exploring scouts methods within a single DFFS run, for example estimating the position of a trial interface using exploring scouts, then, in a second step, firing trial runs to the trial interface to check whether its probability is acceptable.

We have demonstrated the use of both methods, for a simple example of a single particle in a one-dimensional potential (where we showed that the computational efficiency indeed agrees well with our theoretical predictions), and for the more realistic example of the crystallization of Yukawa particles, a computationally intensive system where the nucleation free energy barrier is *a priori* unknown. In the latter case, automatic interface placement led to a large saving in both user and computational time, compared to manual interface placement, even when the manual placement uses some prior knowledge of the shape of the free energy landscape.

The methods presented here should greatly improve the feasibility and computational efficiency of DFFS simulations for computationally expensive systems. Of course, our methods and the approach of Borrero and Escobedo [38] are not incompatible: having placed a set of interfaces automatically using either the trial interface method or the exploring scouts method, one can further optimize their placement iteratively via the method of Borrero and Escobedo, if necessary. For the rare event problem tested here (the crystallization of Yukawa particles), we found that this did not result in any further improvement.

Our focus here has been on automatic interface placement for direct FFS (DFFS) simulations, in which the entire ensemble of trajectories is propagated forward in order parameter space, one interface at a time. In other variants of the FFS method (e.g. branched growth, Rosenbluth-like sampling [20, 22, 28], S-PRES [30] or NS-FFS [31]), transition paths from initial to final state are instead generated in a one-at-a-time fashion. It should be possible to develop modifications of the automatic interface placement methods for these FFS variants: for example applying either the trial interface or exploring scouts method to fix the interfaces during the generation

of the first transition path. The approaches presented here should also be compatible with other interface-based rare event simulation methods such as transition interface sampling [8, 9]. Finally we note that the methods described here could be extended to interfaces that depend on more than one order parameter. For example, in the exploring scouts method, one might track the trajectories of the scouts in two coordinates and set the interfaces to be optimal lines in the space of these coordinates. As well as optimising interface placement, this could also provide a way to adjust the choice of reaction coordinate, on-the-fly during an FFS simulation.

The methods described in this paper have already been implemented in the parallel rare event simulation framework FRESHS [46], which allows the generic use of both FFS and other rare event simulation methods. This framework will soon be publicly available as an open-source package.

Acknowledgments

The authors thank Kevin Stratford, Juho Lintuvuori and Chantal Valeriani for helpful discussions. R.J.A. was funded by a Royal Society University Research Fellowship and by EPSRC under grant EPSRC/EP/I030298/1. A.A. and K.K. were funded by the cluster of excellence “SimTech”, University of Stuttgart.

Appendix A: Optimal flux calculations

Here we describe in more detail our theoretical analysis of the computational efficiency of DFFS, and present our analytical expression for the efficiency as a function of the transition probability. We assume that the transition probability $p_i = p$ is the same for all interfaces. In contrast to the work of Borrero and Escobedo [38], we do not fix the number n of interfaces. Instead, n is determined by the relation

$$n = \frac{\log P_B}{\log p}. \quad (\text{A1})$$

where $P_B = \prod p_i$ is the probability that a trajectory leaving A reaches B before returning to A . Following [21], we define the computational efficiency as

$$\mathcal{E} = \frac{1}{C\mathcal{V}} \quad (\text{A2})$$

where C and \mathcal{V} represent the computational cost of an FFS calculation, and the statistical error (variance) in the resulting rate constant measurement. We use the expressions for C and \mathcal{V} derived in [21] to predict the dependence of \mathcal{E} on the transition probability p .

1. Computational cost

The computational cost of a DFFS calculation, in simulation steps, is approximated in [21] by

$$C \approx N_0 R + M \sum_{i=1}^{n-1} C_i \quad (\text{A3})$$

where N_0 is the number of configurations stored at λ_A , R is the cost of generating each of these configurations, M is the number of trials per interface (note we have assumed this to be constant) and C_i is the average cost of firing a trial run from interface i . Note that Eq.(A3) describes the total cost of the FFS run rather than the cost per configuration at λ_A , as in ref. [21]. Simplifying somewhat the calculation in [21], we assume that the cost of a trial run is linearly proportional to the number of interfaces that it crosses, with proportionality constant S/n (since the spacing between interfaces is inversely proportional to n ; thus S is the cost of a trajectory from A to B). Thus a trial run from λ_i to λ_{i+1} has cost S/n while a run from λ_i to λ_A has cost iS/n . Taking into account the relative probabilities of these outcomes gives

$$C_i \approx \frac{S}{n} [p + i(1-p)] \quad (\text{A4})$$

resulting in the following expression for the cost:

$$\begin{aligned} C &\approx N_0 R + \frac{SM}{n} \sum_{i=1}^{n-1} [p + i(1-p)] \\ &= N_0 \left(R + \frac{Sk}{2n} [2p(n-1) + n(n-1)(1-p)] \right) \end{aligned} \quad (\text{A5})$$

where $k \equiv M/N_0$. The first result in Eq.(A5) is identical to Eq.(A3) in the main text. Substituting in the expression for n in terms of P_B we obtain an expression for the cost in terms of p and P_B :

$$\begin{aligned} C &= \frac{N_0}{2 \log p \log P_B} \cdot [2R \log P_B \log p \\ &\quad + Sk(3p \log P_B \log p + \log P_B^2 \\ &\quad - p \log P_B^2 - 2 \log p^2 - \log P_B \log p)]. \end{aligned} \quad (\text{A6})$$

2. Statistical error

The statistical error – i.e. the variance in a calculation of the rate constant by DFFS, is approximated as in [21], by

$$\mathcal{V} \approx \sum_{i=1}^{n-1} \frac{(1-p_i)}{p_i M_i} \quad (\text{A7})$$

Setting $p_i = p$ and $M_i = M$ we obtain

$$\mathcal{V} = \frac{1}{Mp} (n-1)(1-p) = \frac{(1-p)}{N_0 k p} \left(\frac{\log P_B}{\log p} - 1 \right). \quad (\text{A8})$$

3. Efficiency

Bringing together Eqs.(A2), (A6) and (A8), we obtain the following expression for the computational efficiency in terms of p and P_B :

$$\mathcal{E} = (2kp \log P_B \log p^2) \cdot [(p-1)(\log P_B - \log p) \cdot (\log P_B \log p (Sk(1-3p) - 2R) + Sk \log P_B^2 (p-1) + 2Skp \log p^2)]^{-1} \quad (\text{A9})$$

Expression (A9) was used to generate the data shown in Figure 2.

-
- [1] G. M. Torrie and J. P. Valleau, Chem. Phys. Lett. **28**, 578 (1974).
 - [2] D. Frenkel and B. Smit, *Understanding Molecular Simulation. From Algorithms to Applications* (Academic Press, Boston, 2002), 2nd ed.
 - [3] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).
 - [4] C. H. Bennett, in *Algorithms for Chemical Computations, ACS Symposium, Series No.46*, edited by R.Christofferson (American Chemical Society, Washington, D.C., 1977).
 - [5] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
 - [6] C. Dellago, P. G. Bolhuis, and P. L. Geissler, Adv. Chem. Phys. **123**, 1 (2002).
 - [7] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
 - [8] T. S. van Erp, D. Moroni, and P. G. Bolhuis, J. Chem. Phys. **118**, 7762 (2003).
 - [9] T. S. van Erp and P. G. Bolhuis, J. Comp. Phys. **205**, 157 (2005).
 - [10] T. S. van Erp, Comp. Phys. Commun. **179**, 34 (2008).
 - [11] A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
 - [12] A. M. A. West, R. Elber, and D. Shalloway, J. Chem. Phys. **126**, 145104 (2007).
 - [13] E. Vanden-Eijnden, M. Venturoli, G. Ciccotti, and R. Elber, J. Chem. Phys. **129**, 174102 (2008).
 - [14] G. Henkelman, B. P. Uberuaga, and H. Jonsson, J. Chem. Phys. **113**, 9901 (2000).
 - [15] G. Henkelman and H. Jonsson, J. Chem. Phys. **113**, 9978 (2000).
 - [16] W. E, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
 - [17] W. E, W. Ren, and E. Vanden-Eijnden, J. Phys. Chem. B **109**, 6688 (2005).
 - [18] G. A. Huber and S. Kim, Biophys. J. **70**, 97 (1996).
 - [19] R. J. Allen, P. B. Warren, and P. R. ten Wolde, Phys. Rev. Lett. **94**, 018104 (2005).
 - [20] R. J. Allen, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. **124**, 024102 (2006).
 - [21] R. J. Allen, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. **124**, 194111 (2006).
 - [22] R. J. Allen, C. Valeriani, and P. R. ten Wolde, Journal of Physics: Condensed Matter **21**, 463102 (2009).
 - [23] C. Valeriani, R. J. Allen, M. J. Morelli, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. **127**, 114109 (2007).
 - [24] E. E. Borrero and F. A. Escobedo, J. Phys. Chem. B **113**, 6434 (2009).
 - [25] E. E. Borrero and F. A. Escobedo, J. Chem. Phys. **127**, 164101 (2007).
 - [26] A. Dickson and A. R. Dinner, Annu. Rev. Phys. Chem. **61**, 441 (2010).
 - [27] A. Warmflash, P. Bhimalapuram, and A. Dinner, J. Chem. Phys. **127**, 154112 (2007).
 - [28] F. A. Escobedo, E. E. Borrero, and J. C. Araque, Journal of Physics: Condensed Matter **21**(33), 333101 (2009).
 - [29] C. Valeriani, Ph.D. thesis, FOM AMOLF (2007).
 - [30] J. T. Berryman and T. Schilling, J. Chem. Phys. **133**, 244101 (2010).
 - [31] N. B. Becker, R. J. Allen, and P. R. ten Wolde, J. Chem. Phys. **136**, 174118 (2012).
 - [32] R. P. Sear, J. Chem. Phys. **128**, 214513 (2008).
 - [33] T. S. van Erp, Adv. Chem. Phys. **151**, 27 (2012).
 - [34] L. Fillion, M. Hermes R. Ni, and M. Dijkstra, J. Chem. Phys. **133**, 244115 (2010).
 - [35] C. Valeriani, E. Sanz, and D. Frenkel, J. Chem. Phys. **122**, 194501 (2005).
 - [36] C. Velez-Vega, E. E. Borrero, and F. A. Escobedo, J. Chem. Phys. **133**, 105103 (2010).
 - [37] E. E. Borrero, and F. A. Escobedo, J. Chem. Phys. **125**, 164904 (2006).
 - [38] E. E. Borrero and F. A. Escobedo, J. Chem. Phys. **129**, 024115 (2008).
 - [39] E. Borrero, M. Weinwurm, and C. Dellago, The Journal of chemical physics **134**, 244118 (2011).
 - [40] F. Cérou and A. Guyader, Stochastic Analysis and Applications **25**, 417 (2007).
 - [41] J. D. Weeks, D. Chandler, and H. C. Andersen, J. Chem. Phys. **54**, 5237 (1971).
 - [42] S. Auer and D. Frenkel, Journal of Physics: Condensed Matter **14**, 7667 (2002).
 - [43] E. Sanz, C. Valeriani, D. Frenkel, and M. Dijkstra, Phys. Rev. Lett. **99**, 055501 (2007).
 - [44] F. E. Azhar, M. Baus, J.-P. Ryckaert, and E. J. Meijer, J. Chem. Phys. **112**, 5121 (2000).
 - [45] H.-J. Limbach, A. Arnold, B. A. Mann, and C. Holm, Comput. Phys. Commun. **174** **9**, 704 (2006).
 - [46] K. Kratzer, J. T. Berryman, A. Taudt, and A. Arnold (in preparation).
 - [47] W. Lechner and C. Dellago, J. Chem. Phys. **129**, 114707 (2008).
 - [48] K. Kratzer and A. Arnold (in preparation).
 - [49] Note that if the system enters the final state during this run it should be returned to the initial state and re-equilibrated. It is also important to note that correct computation of k_{AB} depends on good sampling of the configurations at λ_0 : to avoid correlation between configurations arising from rapid recrossings of λ_0 it is often best to store configurations every m crossings (with $m \approx 10$) rather than at every crossing – see ref [22] for

further practical details.